Learning from Measurements in Crowdsourcing Models: Inferring Ground Truth from Diverse Annotation Types

Paul Felt
IBM Watson*
plfelt@us.ibm.com

Eric K. Ringger
Zillow*
ericri@zillow.com

Jordan Boyd-Graber
Computer Science, iSchool, UMIACS,
and Language Science, University of Maryland
jbg@umiacs.umd.edu

Kevin Seppi
Dept. of Computer Science
Brigham Young University
kseppi@byu.edu

Abstract

Annotated corpora enable supervised machine learning and data analysis. To reduce the cost of manual annotation, tasks are often assigned to internet workers whose judgments are reconciled by crowdsourcing models. We approach the problem of crowdsourcing using a framework for learning from rich prior knowledge, and we identify a family of crowdsourcing models with the novel ability to combine annotations with differing structures: e.g., document labels and word labels. Annotator judgments are given in the form of the predicted expected value of measurement functions computed over annotations and the data, unifying annotation models. Our model, a specific instance of this framework, compares favorably with previous work. Furthermore, it enables active sample selection, jointly selecting annotator, data item, and annotation structure to reduce annotation effort.

Annotierte Korpora ermöglichen überwachtes maschinelles Lernen und Datenanalyse. Um die Kosten für manuelle Annotationen zu vermeiden, werden Aufgaben häufig Internetarbeitern zugewiesen, deren Urteile durch Crowdsourcing-Modelle abgeglichen werden. Wir nähern uns dem Problem des Crowdsourcings, indem wir einen Rahmen für das Lernen aus reichem Vorwissen vorschlagen, und wir bestimmen eine Familie von Crowdsourcing-Modellen mit der Fähigkeit, Annotationen mit unterschiedlichen Strukturen zu kombinieren: z.B., Dokumentbezeichnungen und Wortbezeichnungen. Bewertungen werden in Form des vorhergesagten erwarteten Werts von Messfunktionen (measurement functions) gegeben, die über Annotationen und die Daten berechnet werden. Darin werden die vorherige Annotationsmodelle vereinheitlicht. Unser Modell, eine spezifische Instanz dieses Rahmens, schneidet im Vergleich zu früheren Arbeiten positiv ab. Darüber hinaus ermöglicht es die aktive Stichprobenauswahl, indem Kommentator, Datenelement, und Annotationsstruktur gemeinsam ausgewählt werden, um den Annotationskosten zu reduzieren.

1 Introduction

Supervised machine learning is data hungry: new approaches require massive training sets. These training sets can come from inexpensive crowdsourcing platforms, but consistency is often sacrificed for speed and thrift. Sophisticated models (Surowiecki, 2005; Snow et al., 2008; Jurgens, 2013) can overcome the intrinsic annotation noise by reconciling redundant annotation and predicting true labels by modeling the error patterns associated with individual labels, documents, or annotators.

These models have typically assumed that we collect document-level *labels* and nothing else from annotators. But crowd workers could provide other valuable information. For example, if we wanted to predict the sentiment of documents about the weather (Figure 1), intuition says that words like "sunny",

* This work was completed while the first and second authors were at Brigham Young University.

- *. . . Sunshine: 60s and partly sunny? OK!*
- *. . . shopping, sunshine, margaritas, happiness!*
- *Going to the zoo . . . the weather is perfect.*
- *Another rainy day! Blah.*
- *Cold and rainy here in Boston, Wish I was in ATL*
- *Damn its hot out. Even when not working . . .*
- *Horrible wet morning . . . bring back the sunshine!*

Figure 1: Example tweets from the "Weather Sentiment" dataset available at `http://www.crowdflower.com/data-for-everyone`. Annotators can annotate both documents and *measurements* (e.g., whether a word appears in a document) to label sentiment.

"sunshine", and "perfect" will often appear in positive tweets, and words like "blah", "cold", and "hot" will tend to be more negative. This information is often complementary to labels attached to documents.

However, existing crowdsourcing models cannot simultaneously model multiple kinds of information. This paper introduces a framework and model to combine such diverse information from crowd workers, to measure and model individual annotator errors, and to form consensus final predictions.

We build on the measurements annotation framework (Section 2), which has previously only been applied to the scenario of a single, trusted annotator. Section 3 discusses the application of the framework to crowdsourcing scenarios with multiple untrusted annotators and show that our annotator-aware measurement framework subsumes existing crowdsourcing models. In Section 4 we develop inference for a crowdsourcing measurements model and show that it captures per-annotator variance on both simulated and crowdsourced data in Section 5. We then further extend the model for active query strategies, asking crowd workers for mixed annotations (Section 6).

2 Single-Annotator Measurements

Liang et al. (2009) introduce a supervised learning framework that uses more than the annotation of a document. This section reviews the intuition and notation of this framework that we extend to multiple annotators in the following sections.

The key intuition of the measurement framework is that annotators often have insights that generalize beyond a single document x and label y. For example, in a sentiment labeling task, an annotator could supply the clue that the word "lackluster" often appears in documents with negative sentiment.

These insights are encoded through functions called *measurement features*: $\sigma_k : \mathcal{X}, \mathcal{Y} \mapsto \mathcal{R}$. The measurement feature σ_k tests whether property k holds for the pair x, y, where x is a data item such as a sentence or document and y is its (possibly structured) label. The index k encodes everything needed by a measurement feature to fire only in an extremely specific situation. Measurement features are more specific and correspondingly more sparse than traditional NLP features.

Measurement features can encode traditional supervised labeling. For example, in sentiment classification some measurement feature $\sigma_{k'}$ might encode the fact that Document 343 has a positive label by being zero except when x exactly matches Document 343 and y is positive: $\sigma_{k'} = \mathbb{1}(x = x_{343}, y = Positive)$.

But introducing measurement functions also allows more flexibility. Returning to our "lackluster" example, we can include a function that is one if that word is in the example and the label is negative: $\sigma_{k''} = \mathbb{1}(\text{"lackluster"} \in x, y = Negative)$.

Thus measurement features map the data into an extremely sparse and high-dimensional (partially) observed space. Moreover, measurement features can span multiple documents, so each measurement feature is summed over the dataset $\sigma_k(x, y) = \sum_i \sigma_k(x_i, y_i)$. The learning from measurements framework defines K measurement features, one for every possible labeling. Table 1 provides examples of properties that measurements can encode.

The measurements framework treats observed measurement values τ as the result of measurement noise ϵ applied to the result of measurement features σ.

Measurement Type	Observed $\sigma_k(x, y)$	Partially Observed $\sum_i E_{q(y_i)}[\sigma_k(x, y)]$	Maximum $\max(\sigma_k)$
Document Label	$\mathbb{1}(x = x_m, y = c)$	$q(y_m = c)$	1
Word Label	$\mathbb{1}(f(x) = 1, y = c)$	$\sum_{i \in f(X)} q(y_i = c)$	$\sum_i \mathbb{1}(f(x_i) = 1)$
Label Proportion	$\mathbb{1}(y = c)$	$\sum_i q(y_i = c)$	N

Table 1: The measurement paradigm reformulates the direct labels of traditional supervised learning as indirect measurement features σ and their expected values. If we could directly observe class labels c then we would compute σ (*Observed* column). Since we only have indirect annotation evidence we must learn via the expected values in the *Partially Observed* column. $\mathbb{1}(\cdot)$ is an indicator function. Expected values are defined with respect to the approximate model q (defined in Section 4). All table values remain the same when dealing with annotator-indexed measurements σ_{jk}, although σ_{jk} additionally encodes annotator identity j as well as an implicit annotation value k.

Figure 2 illustrates the measurement framework's generative[1] story:

1. Draw parameter vector θ.
2. Draw measurement noise prior γ.
3. For $i \in N$ instances:
   (a) Draw label y_i from conditional exponential model family $p(y_i \mid x_i, \theta)$.
4. For $k \in K$ measurements:
   (a) Draw measurement noise ϵ_k from $p(\epsilon_k \mid \gamma)$.
   (b) Draw measurement τ_k from $p(\tau_k \mid \sum_i \sigma_k(x_i, y_i), \epsilon_k)$.

Although document labels y are part of the hypothetical generative story according to this model, at inference time y is always unobserved (Figure 2). Like many crowdsourcing models, while the true labels y cannot be observed directly, some byproduct τ of label y can provide evidence for inferring y.

While measurement noise ϵ_k is a part of the generative model, previous implementations of measurement models ignore this component, effectively assuming that all measurements have the same noise. Prior work ignored these considerations because traditional supervised learning training sets lack information about annotators to effectively model per-annotator or per-measurement noise.

In the next section, we correct this omission by extending the measurement model to specifically model not just the measurement noise model but to also estimate the *source* of these errors.

3 Connecting Measurements and Crowdsourcing

This section applies the measurements framework (Section 2) to crowdsourcing scenarios with multiple untrusted annotators and shows the connection to existing crowdsourcing frameworks. The adapted framework (Figure 2) requires two changes to the original measurements framework. First, each measurement k is replicated for each annotator j. This re-indexing of k accommodates annotator-specific parameters that can encode the expertise or focus of particular annotators in a crowdsourcing framework. The generative process is unchanged except for the final step:

4. For $j \in J$ annotators:
   (a) For $k \in K$ measurements:
      i. Draw measurement noise ϵ_{jk} from $p(\epsilon_{jk} \mid \gamma)$.
      ii. Draw measurement τ_{jk} from $p(\tau_{jk} \mid \sum_i \sigma_{jk}(x_i, y_i), \epsilon_{jk})$.

In addition to drawing per-annotator measurements, we also add a hierarchical prior γ over noise parameters (omitted here, as we consider multiple forms of the prior later): this induces parameter tying among measurement noise distributions. For example, this tied parameter can encode trust in annotator j by tying all ϵ_{jk} for annotator j and be left with a single noise parameter ϵ_j per annotator by imposing a prior where $\forall j \exists k, k'(\epsilon_{jk} \neq \epsilon_{jk'}) \implies p(\epsilon|\gamma) = 0$.

The measurements framework leaves the structure of y, the conditional exponential family used to

[1]The measurements framework omits the extra plate indexed by j. It will be addressed in the next section.

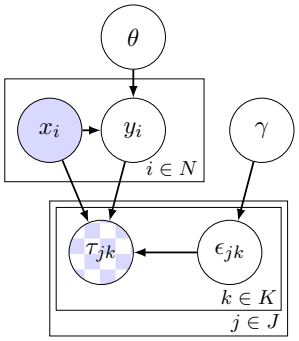


Figure 2: Plate diagram of the measurements framework (Liang et al., 2009) adapted for multiple untrusted annotators. Shaded nodes have observed values. Partially shaded nodes have some observed values. This model subsumes other crowdsourcing models with naïve measurements and extends them to varied annotation environments with richer measurement functions.

model $p(y \mid x, \theta)$, and definitions of the distributions $p(\theta)$ and $p(\epsilon_{jk})$ unspecified. This paper situates the measurements framework into existing crowdsourcing models. To see the connection, let y_i take on discrete class values, let $p(y_i \mid x_i, \theta) = p(y_i \mid \theta)$ be a data-agnostic multinomial distribution, and let each observed annotation define a document label measurement σ_{jk} (Table 1) where k encodes a specific instance indexed by i, annotation value c', and label value c. Finally, let noise parameters ϵ_j for annotator j be tied to a confusion matrix with Dirichlet-distributed rows such that ϵ_{jk} selects the value at cell (c, c'), encoding how likely j is to produce annotation c' when shown a document whose true label is c. These settings recover the traditional item-response crowdsourcing model (Dawid and Skene, 1979). Using the same settings but defining $p(y \mid x, \theta) \propto exp[\theta^T f(x, y)]$, recovers a popular data-conditional crowdsourcing model (Raykar et al., 2010; Yan et al., 2014; Felt et al., 2015b). However, existing crowdsourcing models lack the representational richness of the measurements framework; we address this lacuna in the next section.

4 Per-annotator Normal Measurement Model for Classification

Having shown how the annotator measurements framework can capture existing crowdsourcing models, this section presents a novel crowdsourcing model that instantiates the richness of the measurements framework that we use in Sections 5–6. For brevity, we refer to this model as PAN (per-annotator normal) measurement model.[2] The generative story is:

1. Draw a stochastic vector θ over C classes from a symmetric Dirichlet Dir(δ)
2. For $i \in N$ documents, draw label y_i from categorical Mult(θ).
3. For $j \in J$ annotators:
   (a) draw measurement noise ϵ_j from inverse gamma IG(α, β).
   (b) For $k \in K$, draw measurement τ_{jk} from a normal Norm($\sum_i \sigma_{jk}(x_i, y_i), \epsilon_j$))

4.1 Learning by Variational Inference

The PAN model's conditional log joint distribution is

$$\log p(\theta, y, \tau \mid x) = \tag{1}$$
$$-\log \text{Beta}(\delta) + J\alpha \log \beta - J \log \Gamma(\alpha) - \sum_j \frac{K_j}{2} \log(2\pi) + \sum_c (\delta + n_c - 1) \log \theta_c$$
$$+ \sum_j (-(\alpha + \frac{K_j}{2}) - 1) \log \epsilon_j - \left(\frac{\beta + \frac{1}{2} \sum_k (\tau_{jk} - \sum_i \sigma_{jk}(x_i, y_i))^2}{\epsilon_j} \right)$$

[2]Data and code available at `https://github.com/BYU-NLP-Lab/Experiments`.

where $\beta\cdot$ is the multivariate beta function and K_j is the number of measurements values observed by annotator j. We use mean-field variational inference as an efficient approximation to full likelihood maximization by assuming a fully factorized approximate model:

$$q(\theta, y \mid \nu) = q(\theta \mid \nu^{(\delta)}) \prod_i q(y_i \mid \nu_i^{(y)}) \prod_j q(\epsilon_j \mid \nu_j^{(\alpha)}, \nu_j^{(\beta)})$$

and then minimizing Kullback-Leibler divergence $\mathrm{KL}(q \,\|\, p)$ via coordinate ascent by iteratively updating the variational parameters ν. That is, $q(\theta \mid \nu^{(\delta)}) \sim \mathrm{Dir}(\nu^{(\delta)})$ where

$$\nu_c^{(\delta)} = \delta + \sum_i \nu_{y_i,c}^{(\delta)} \ . \tag{2}$$

Similarly, $q(\epsilon_j|\nu_j^{(\alpha)}, \nu_j^{(\beta)}) \sim \mathrm{IG}(\nu^{(\alpha)}, \nu^{(\beta)})$ where

$$\nu_j^{(\alpha)} = \alpha + \frac{K_j}{2}, \quad \nu_j^{(\beta)} = \beta + \frac{1}{2} \sum_{k \in S(j)} \mathbb{E}_{q(y_i)} \left[\left(\sum_i \sigma_{jk}(x_i, y_i) \right)^2 \right] \tag{3}$$

where $S(j)$ is the set of measurements provided by annotator j. Although the term $\mathbb{E}_{q(y_i)} \left[(\sum_i \sigma_{jk}(x_i, y_i))^2 \right]$ appears intractable, we can simplify it by introducing terms to complete the square.

$$\mathbb{E}_{q(y_i)} \left[\left(\sum_i \sigma_{jk}(x_i, y_i) \right)^2 \right] = \tag{4}$$
$$\sum_i \left(\mathbb{E}_{q(y_i)} \left[\sigma_{jk}(x_i, y_i) \right] \right)^2 - \sum_i \mathbb{E}_{q(y_i)} \left[\sigma_{jk}(x_i, y_i) \right]^2 + \mathbb{E}_{q(y_i)} \left[\sigma_{jk}(x_i, y_i)^2 \right] .$$

Finally, $q(y_i \mid \nu_i^{(y)}) \sim \mathrm{Mult}(\nu_i^{(y)})$. We update $\nu_i^{(y)}$ by evaluating $\log \nu_i^{(y)} + const$ for each setting of y_i and then exponentiating and normalizing the resulting vector.

$$\log \nu_i^{(y)} = \psi(\nu_{y_i}^{(\delta)}) - \psi(\sum_{y_i} \nu_{y_i}^{(\delta)}) + const + \sum_j \frac{\nu_j^{(\alpha)}}{2\nu_j^{(\beta)}} \Bigg(\sum_{k \in S(i,j)} 2\tau_{jk}\sigma_{jk}(x_i, y_i) \tag{5}$$
$$- \sigma_{jk}(x_i, y_i)^2 - 2\sigma_{jk}(x_i, y_i) \sum_{i' \neq i} \mathbb{E}_{q(y_{i'})} \left[\sigma_{jk}(x_{i'}, y_{i'}) \right] \Bigg)$$

where $S(i, j)$ is the set of measurements provided by annotator j that relate to instance i. More formally, $S(i, j)$ is the set of measurements k where there is some setting of y_i that makes the measurement feature $\sigma_{jk}(x_i, y_i)$ evaluate to a non-zero value.

To get a taste of inference in PAN, consider a simple concrete scenario where four people are labeling the sentiment of a tweet: "Wishing good weather were here again". First Alice labels the tweet positive. Lacking any other information, PAN accepts that for now, calling $\nu_0^{(y)} = [0.75, 0.25]$ and assigning Alice moderate trust in the form of $\mathrm{IG}(\nu_{alice}^{(\alpha)} = 1.6, \nu_{alice}^{(\beta)} = 1.22)$ with a mean error rate of 2.03. Next Bob labels the tweet negative. Since it has no reason to trust Bob more than Alice, the model splits the class vote evenly $\nu_0^{(y)} = [0.5, 0.5]$ and downgrades its trust in both Alice and Bob because of the conflict, assigning them both error rate 2.25. Next Carol labels the word "good" as being positive. Since the word "good" is in our document this tips the balance in favor of positive: $\nu_0^{(y)} = [0.8, 0.2]$. Notice the positive balance is more than $2/3$ since Alice and Carol are now aligned with the model's belief about the true labels and their error rate is upgraded to 2.0, while dissenting Bob's error rate is downgraded to 2.5. Finally, Dave, who happens to be highly trusted *a priori* (perhaps because of admin status or previous good work) comes along and labels the word "wishing" as negative. Since Dave has clout the document swings negative $\nu_0^{(y)} = [0.2, 0.8]$, and Dave and Bob are now aligned with the truth, while Carol and Alice are not. Dave and Bob's error rates go to 0.56 and 2.0, respectively, while Alice and Carol's degrade to 2.5.

4.2 Implementation Considerations

The updates in the previous section contain terms like $\sum_j \sum_k \sum_i \mathbb{E}_{q(y_i)} [\sigma_{jk}(x_i, y_i)]$. Despite the apparent expense of these terms, many of these sums are very sparse and may be computed efficiently. In addition, these values may be cached and incrementally updated as necessary to reduce computational expense.

The scale of measurement features can confuse both users and the algorithm. In Equation 1, observed measurement values τ are compared with the value of measurement features summed over the dataset $\sum_i \sigma(x_i, y_i)$. This latter quantity is bounded by a different range for each measurement feature (see Table 1). However, humans may prefer to give measurement values between 0 and 1, where 1 means "this happens as often as possible." Such values must be scaled by $\max(\sigma_{jk})$.

A related point is that each measurement type is defined on a different scale, but the PAN model fits a common variance to all types. For this to make sense, it is necessary to re-scale each measurement to a common range before running inference. For all experiments reported in this paper, we choose the range $[0 \dots 1]$. Concretely, substitute $\frac{\sigma_{jk}(x_i,y_i)}{\max(\sigma_{jk})}$ for $\sigma_{jk}(x_i, y_i)$, $\frac{E_{q(y_i)}[\sigma_{jk}(x_i,y_i)]}{\max(\sigma_{jk})}$ for $E_{q(y_i)}[\sigma_{jk}(x_i, y_i)]$, and $\frac{\tau_{jk}}{\max(\sigma_{jk})}$ for τ_{jk}.

5 Experiments

This section explores the performance of PAN in rich annotation scenarios. Our goal here is not to establish the PAN model as the state of the art but rather to assess and validate the utility of incorporating diverse annotation types.

5.1 Baselines

We choose two common crowdsourcing baselines which are widely compared against. Despite their simplicity, previous work in establishing benchmark crowdsourcing tasks indicates that these baselines are surprisingly competitive with more sophisticated methods (Sheshadri and Lease, 2013).

Majority Vote (MV) chooses the label with the largest number of votes for each item, ignoring annotator identify. Ties are broken randomly. Although simple, majority vote is widely used and surprisingly successful across a variety of tasks.

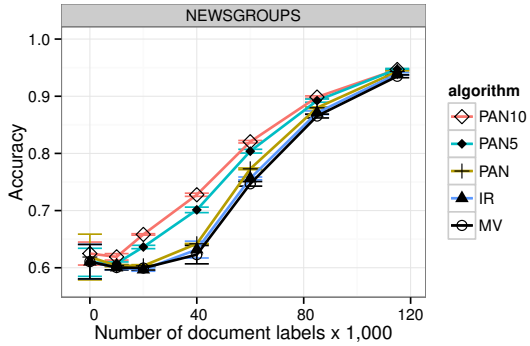
Dawid & Skene (IR). The *item response model* proposed by Dawid and Skene (1979) models annotator error over discrete responses via per-annotator confusion matrices. Much subsequent work uses the same basic structure, making it a good point of comparison. We use a Bayesian version of this model with variational inference adapted from Felt et al. (2015b). The general learning from measurements crowdsourcing framework generalizes this model and many of its extensions (Section 3).

5.2 Simulated Data

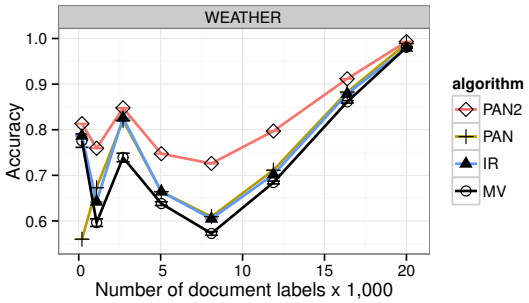
We first generate confidence in our implementation by running on the well-known *20 News Groups* dataset with simulated annotator measurements consistent with PAN's Gaussian noise assumptions. We simulate three different measurements: document labels, word labels, and estimated label proportions. Document label judgments are simulated by corrupting the known true document labels via confusion matrices for five annotators with 50%, 55%, 60%, 65%, and 70% accuracy rates, using the following process:

1. Choose a document randomly without replacement. If none are left, begin again.
2. Choose a simulated annotator randomly with replacement and annotate the document according to that annotator's accuracy.
3. Stop when each of the 20,000 documents has approximately six label annotations.

Word label judgments are simulated by choosing a word w and document label y uniformly at random, and then calculating the empirical rate of seeing documents with label y given that they contain word w. We add Gaussian noise proportional to the total number of documents containing word w and inversely proportional to the accuracy of the annotator. Finally, we manually specify twenty label proportion judgments to encode our *a priori* belief that classes appear roughly the same number of times in the data.



(a) Simulated annotations over the 20 newsgroups corpus. The corpus is annotated until each document has approximately six document label annotations. Performance is competitive with methods with only document labels. However, as we add additional word labels—5,000 (*PAN5*), and 10,000 (*PAN10*)—via the measurements framework, performance increases further.



(b) Real annotations over the weather dataset. PAN is with (PAN2) and without (PAN) access to 2,266 labeled words from CrowdFlower. Labeled words substantially improve PAN's predictions.

Figure 3: Inferred labeled accuracy of Majority vote (MV), the item-response (IR) model, and the per-annotator normal (PAN) measurement model.

In practice, such prior knowledge about approximate label proportions may be available depending on how the data was gathered.

5.3 Inferred label accuracy curves

When the annotation process begins, few documents have annotations. Because crowdsourcing models require annotations to make a prediction, we can plot the accuracy of inferred labels only over those documents having at least one annotation. This means that until the dataset is annotated once, the denominator of the accuracy calculation is growing. Thus inferred label accuracy reflects the corpus accuracy over the subset of documents that have at least one label. This can make inferred label accuracy curves look unlike traditional learning curves, especially when annotation order is not controlled. In real annotation projects, some documents might be annotated multiple times before other documents receive any annotations, resulting in potential accuracy dips as the denominator changes.

PAN is competitive with baselines using only simulated document labels, and it benefits from additional simulated word labels (Figure 3a). Unsurprisingly, there are diminishing returns from additional labeled words: the difference between 0 and 5,000 labeled words is more dramatic than the difference between $5,000$ and $10,000$.

5.4 Sentiment Classification

The same trends hold with real annotator judgments using the "Weather Sentiment" dataset. In this dataset twenty annotators label 1,000 tweets as either `Positive`, `Negative`, `Neutral`, or `Not weather`. Majority vote labels are then evaluated in the related "Weather Sentiment Evaluated" task where a ten secondary annotators judge whether each consensus label is correct or not. For 724 of the tweets, at least nine secondary annotators agree that the consensus majority vote label is correct. We use these high confidence tweets as our gold standard.

Document label judgments are already available for this dataset, but no labeled words or label proportion judgments. We paid CrowdFlower workers to generate labeled words by showing them groups of then randomly selected weather tweets and asking them to list words that characterize each class of tweet. Although a more highly trained workforce could have generated and labeled more sophisticated measurements such as regular expressions, labeled words are a first test. Furthermore, we encode an *a priori* belief that each class occurs roughly the same number of times by manually adding four trusted label proportion measurements stating that each class occurs $N/C = 250$ times. Trusted measurements

Algorithm 1: Active measurement selection algorithm for jointly selecting annotator j and annotation type k. The NEXTMEASUREMENT subroutine approximates expected utility.

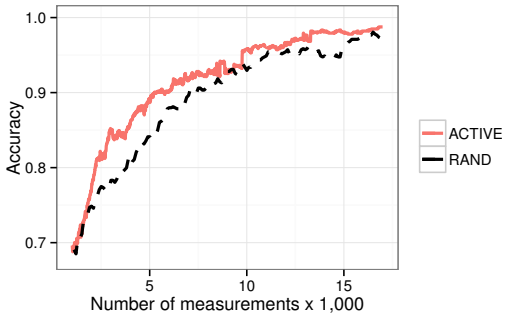


Figure 4: Inferred labeled accuracy of the PAN model selecting measurements randomly (*RAND*) compared with a strategy that selects measurements and annotators actively (*ACTIVE*).

are expressed in this framework by authoring measurements under the id of an artificial annotator who is assigned a strong prior distribution of low measurement noise.

From the workers we gather 864 word lists containing 2,482 individual words and covering a vocabulary of 995 unique words. Of the 2,482 word labels, 216 did not match any words in the corpus (we allowed users the freedom to draw on their own background knowledge as well as words of the corpus that they were shown). Interestingly, a brief manual examination of the word lists did not uncover any abusive behavior, although several annotators clearly wanted to match phrases rather than just words. Although the model would permit the use of labeled phrases, our Crowdflower task was not designed to distinguish the two cases, so for now we treat all words as individual judgments. We match word labels to document words after normalizing both by removing punctuation, converting to lower case, applying a Porter stemmer, and removing words from the MALLET stopword list (McCallum, 2002).

Labeled words improve PAN's accuracy. In Figure 3b, the PAN2 line uses $2k$ labeled words and beats vanilla PAN substantially. And while labeled words are not free, it took only 864 judgments to generate our labeled word set, which works out to about $9 at $0.01 per judgment; whereas the 20,000 judgments comprising the document labels would have cost $200 at $0.01 per judgment.

6 Active Measurement Selection

The learning curves in Section 5 assume that annotations were obtained in some arbitrary order, either randomly (Section 5.2) or else ordered by timestamp (Section 5.4). Active learning minimizes annotation costs by obtaining annotations in an order that maximizes their utility. Previous work in active learning either assumes a single, infallible annotator (Settles, 2010; Liang et al., 2009), or else allows multiple annotators but assumes a single kind of annotation (Donmez and Carbonell, 2008; Haertel, 2013; Yan et al., 2011; Nguyen et al., 2015). This paper presents the first active learning results in a setting with multiple annotators and diverse annotation types, jointly selecting annotator, document, and annotation type.

We adapt the active measurement selection algorithm of Liang et al. (2009) to fit the crowdsourcing scenario in Algorithm 1. At each step in the MEASUREMENTSELECTION subroutine, we have a set of observed measurement labels τ_0 and wish to observe a new measurement feature $\sigma_{j,k}$ encoding both the annotator j and the annotation type k (including the document to be annotated for document-centric measurement features).

The NEXTMEASUREMENT subroutine in Algorithm 1 contains our selection criteria, and can be understood as approximating expected net utility:

$$U(\sigma_{jk}) = \mathbb{E}_{p(\tau_{jk} \mid \tau_0, X)} \left[R(\sigma_{jk}, \tau_{jk}) - C(\sigma_{jk}) \right]$$

where $R(\sigma_{jk}, \tau_{jk})$ is the expected reward for obtaining judgment value τ_{jk} for measurement feature σ_{jk}, and $C(\sigma_{jk})$ is the expected cost of obtaining that judgment. To make this computation tractable, we introduce a number of approximations. Lines 11–15 of Algorithm 1 compute the expectation over τ_{jk} using stochastic integration. For PAN, we approximate sampling from the posterior $p(\tau_{jk} \mid \tau_0, X)$ by parameterizing the Normal distribution of the original PAN model with the mean value of our variational parameters: $p(\tau_{jk} \mid \tau_0, q_0) = \text{Norm}(\sum_i \sigma_{jk}(x_i, y_i), \frac{\nu^{(\beta)}}{\nu^{(\alpha)}-1})$.

The expected reward function R should reflect our expected satisfaction at having observed τ_{jk}. Ideally, we would be able to compute model improvement directly by comparing true document labels y with our predicted labels $\hat{y}$ after adding the new observation: $\mathbb{E}_{p*(x)}\left[\max_{\hat{y}} r(y, \hat{y})\right]$ where $r(y, \hat{y})$ is an internal reward function like label accuracy and $p^*(x)$ is the empirical distribution. In reality the true values y are unobservable, but we can expect over them using our posterior approximation $\tilde{q}$; thus $R_{\tilde{q}} = \mathbb{E}_{p*(x)}\left[\max_{\hat{y}} \mathbb{E}_{\tilde{q}(y)}\left[r(y, \hat{y})\right]\right]$. With a label accuracy reward function the expected reward simplifies to $R_{\tilde{q}} = \sum_i \max_{\hat{y}} q(y_i = \hat{y}) = \sum_i \max_{\hat{y}} \nu_{i\hat{y}}^{(y)}$. For simplicity, we set the cost function $C_{\tilde{q}}$ to a constant, but leave it in the equations since future work should estimate and use annotation cost.

By default, Algorithm 1 jointly selects an annotator j and measurement k, but it could be used in other ways. Haertel et al. (2010) argue that in realistic scenarios the active learning algorithm typically cannot control when annotators are available but rather must respond to annotator requests for work. Algorithm 1 can return the best measurement k given annotator $\hat{j}$ by computing $\text{argmax}_k \mu_{\tau_{\hat{j}k}}$. Similarly, one can calculate the best annotator j for a desired measurement $\hat{k}$ as $\text{argmax}_j \mu_{\tau_{j\hat{k}}}$.

Unfortunately, Algorithm 1 is computationally expensive. Prior to selection, each candidate measurement must be considered and a model retrained using t sampled candidate measurements. However, by applying a number of additional approximations we can run on the weather sentiment dataset from Section 5.4 with over 22,000 candidate measurements. We set the number of samples to three, and models trained in the inner loop (line 14 of Algorithm 1) are initialized using q_0 and then trained only one additional iteration. More importantly, we score only twenty-five randomly selected candidates at each round and select batches of the ten most promising measurements at each round.

We compare random and active selection of measurements (Algorithm 1) from the sentiment classification experiment in Section 5.4 in Figure 4. The active measurement selection improves over a random baseline.

7 Additional Related Work

Other supervised learning frameworks that incorporate rich prior knowledge include constraint-driven learning based on integer linear programming (Chang et al., 2008), generalized expectation criteria (Druck et al., 2008), and the posterior regularization (Ganchev et al., 2010). Ganchev et al. (2010) explain each of these three frameworks can be derived as a special case of the learning from measurements framework of Liang et al. (2009) by making particular approximations for the sake of tractability.

Traditional corpus construction assesses inferred label quality using annotator agreement heuristics such as Krippendorff's alpha (Krippendorff, 2012). Passonneau and Carpenter (2014) argue that inference in probabilistic models yields higher quality labels at lower cost, and should be preferred over agreement heuristics. Among crowdsourcing models, Hovy et al. (2013) include Bernoulli switching variables to identify and eliminate malicious contributors (spammers). Raykar and Yu (2012) iteratively run inference and exclude problematic annotators in order to eliminate spammers. Raykar et al. (2010), Yan et al. (2014), and Felt et al. (2015a) model data jointly with labels, allowing patterns in the data to inform inferred labels. Simpson and Roberts (2015) model annotator dynamics, tracking the ways that annotator decision making changes over time in response to factors such as training and fatigue. Welinder et al. (2010) and Whitehill et al. (2009) both model item difficulty, reducing the effect of inherently ambiguous or difficult items on annotator reliability estimates.

Each of these crowdsourcing models focuses on incorporating one or more insights about the annotation process. We leave it to future work to incorporate these insights into crowdsourcing models that learn from measurements, either via measurement noise or via new measurement formulations. For example, item difficulty could be modeled by creating a measurement feature $\sigma_{jk}(x_i = x_{i'})$ for each $x_{i'}$

and assigning a separate measurement noise to each (perhaps with a shared hierarchical prior noise).

8 Conclusion and Future Work

The success of machine learning depends on data, and those data often come from human annotators. Asking the right questions of the right people is an often overlooked challenge of building an effective machine learning system. Extending the measurements framework allows modeling users' quirks and using their insights more effectively.

This paper makes a focused contribution by connecting the measurements framework with crowdsourcing models and using initial experiments to showcase the flexibility and promise of this connection. However, the measurements framework is far more general than this first round of experiments is able to show. One primary direction of follow-on work will be to extend crowdsourcing measurement models to more sophisticated structured prediction applications. Another will be to develop inference for a more robust noise model. For example, Gaussian Processes could be used as measurement noise priors to capture more complex interactions between measurement types and annotators.

As we move toward richer annotations, we also need to consider the implications for human interactions. Modeling the costs of annotations can prevent crowdsourcing from asking difficult, ambiguous, or annoying questions. Potentially interesting measurements may include allowing annotators to report their own reliability, to assess the reliability of other annotators, or to label locations in a semantically meaningful space rather than discrete words or documents.

Acknowledgments This work was supported by the collaborative NSF Grant IIS-1409739 (BYU) and IIS-1409287 (UMD). Boyd-Graber is also supported by NSF grant IIS-1652666. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

Ming-Wei Chang, Lev-Arie Ratinov, Nicholas Rizzolo, and Dan Roth. 2008. Learning and inference with constraints. In *Proc. Conference on Artificial Intelligence (AAAI)*.

Alexander P. Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, pages 20–28.

Pinar Donmez and Jaime G. Carbonell. 2008. Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In *Proc. ACM Conference on Information and Knowledge Management*.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Paul Felt, Eric K. Ringger, Jordan Boyd-Graber, and Kevin Seppi. 2015a. Making the most of crowdsourced document annotations: Confused supervised LDA. In *Proc. Conference on Computational Natural Language Learning (CoNLL)*.

Paul Felt, Eric K. Ringger, Kevin Seppi, and Robbie A. Haertel. 2015b. Early gains matter: A case for preferring generative over discriminative crowdsourcing models. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049.

Robbie A. Haertel, Paul Felt, Eric K. Ringger, and Kevin Seppi. 2010. Parallel active learning: Eliminating wait time with minimal staleness. In *Proc. HLT-NAACL 2010 Workshop on Active Learning for Natural Language Processing*.

Robbie A. Haertel. 2013. *Practical Cost-Conscious Active Learning for Data Annotation in Annotator-Initiated Environments*. Ph.D. thesis, Brigham Young University.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning whom to trust with MACE. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Klaus Krippendorff. 2012. *Content Analysis: An Introduction to its Methodology*. Sage.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning from measurements in exponential families. In *Proc. International Conference on Machine Learning (ICML)*.

Andrew McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

An T. Nguyen, Byron C. Wallace, and Matthew Lease. 2015. Combining crowd and expert labels using decision theoretic active learning. In *Proc. 3rd AAAI Conference on Human Computation (HCOMP)*.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Vikas C. Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *The Journal of Machine Learning Research*, 13:491–518.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *The Journal of Machine Learning Research*, 11:1297–1322.

Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison*.

Aashish Sheshadri and Matthew Lease. 2013. Square: A benchmark for research on computing crowd consensus. In *Proc. Conference on Human Computation and Crowdsourcing (HCOMP)*.

Edwin Simpson and Stephen Roberts. 2015. Bayesian methods for intelligent task assignment in crowdsourcing systems. In *Decision Making: Uncertainty, Imperfection, Deliberation and Scalability*, pages 1–32. Springer.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

James Surowiecki. 2005. *The Wisdom of Crowds*. Random House LLC.

Peter Welinder, Steve Branson, Pietro Perona, and Serge J. Belongie. 2010. The multidimensional wisdom of crowds. In *Proc. Advances in Neural Information Processing Systems (NIPS)*.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L. Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proc. Advances in Neural Information Processing Systems (NIPS)*.

Yan Yan, Glenn M. Fung, Rómer Rosales, and Jennifer G. Dy. 2011. Active learning from crowds. In *Proc. International Conference on Machine Learning (ICML)*.

Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. 2014. Learning from multiple annotators with varying expertise. *Machine Learning*, 95(3):291–327.